Am. J. Hum. Genet. 49:106–111, 1991

Detecting High-Resolution Polymorphisms in Human Coding Loci by Combining PCR and Single-Strand Conformation Polymorphism (SSCP) Analysis

Shirley E. Poduslo,*[,1] Michael Dean,* Ulricke Kolch,* and Stephen J. O'Brien†

*Program Resources, Incorporated/DynCorp, and †Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick Cancer Research and Development Center, Frederick, MD

Summary

A strategy is described that allows the development of polymorphic genetic markers to be characterized in individual genes. Segments of the 3′ untranslated regions are amplified, and polymorphisms are detected by digestion with frequently cutting enzymes and with the detection of single-stranded conformation polymorphisms. This allows these genes, or DNA segments, to be placed on the linkage maps of human chromosomes. Polymorphisms in two genes have been identified using this approach. A *Hae*III polymorphism was detected in the *KIT* proto-oncogene, physically assigned to chromosome 4q11-12. This polymorphism is linked to other chromosome 4p markers and is in linkage disequilibrium with a *Hin*dIII polymorphism previously described at this locus. We have also identified in the insulin-like growth factor1 receptor gene (*IGF1R*) a 2-bp deletion that is present at a frequency of .25 in the Caucasian population. Pedigree analysis with this insertion/deletion polymorphism placed the *IGF1R* gene at the end of the current linkage map of chromosome 15q, consistent with the physical assignment of 15q2526. Thus, polymorphisms in specific genes can be used to relate the physical, genetic, and comparative maps of mammalian genomes and to simplify the testing of candidate genes for human diseases.

Introduction

More than 1,800 DNA probes have been described that detect RFLPs; 1,000 of these have been typed in the CEPH collection of pedigrees (Kidd et al. 1989; Dausset et al. 1990). However the majority of these probes are anonymous DNA markers, not associated with specific genes. Polymorphisms in fewer than 400 human genes have been described, and most of these are not generally useful, because they have a low PIC and/or are detected by infrequently used restriction enzymes.

Received December 12, 1990; revision received February 21, 1991.

Address for correspondence and reprints: Dr. Michael Dean, Building 560, National Cancer Institute, Frederick Cancer Research and Development Center, Frederick, MD 21702-1201.

1. Present address: Department of Neurology, Texas Tech University Health Sciences Center, Lubbock TX 79430.

The development of informative polymorphisms in human coding genes is an important goal for several reasons. First, a critical reason for gene mapping is to resolve the relationship of gene function to development, to disease, and to phenotype. Thus the availability of genotypic variation in coding genes has traditionally formed the basis for genetic investigations. The identification of candidate loci for heritable human disease and phenotypic traits would be facilitated by the availability of a contiguous gene map of polymorphic functional genes. Second, a growing data base of comparative gene mapping in mouse and several other mammalian species is undergoing rapid development. Since comparative genetics requires knowledge of sequence homology, informative extrapolation requires that evolutionarily conserved coding loci be studied (O'Brien and Graves 1990). Ignoring coding loci in human gene mapping would risk loss of this comparative perspective. Third, as genetic analysis pervades all fields of human biology, cumulative genetic information based on health and

biological questions (not necessarily in the purview of the human genome project) will continue to contribute to our understanding of genome organization. Development of polymorphic coding genes (in contrast to anonymous DNA segments) would be a rewarding investment in the stimulation of such investigations.

Traditional methods to identify RFLPs in specific genes involve the hybridization of cDNA or genomic probes to DNA from unrelated individuals which has been digested with a number of restriction enzymes. This approach has been successful in finding a moderately frequent polymorphism, in a commonly used enzyme, approximately a third of the time (M. Dean, unpublished data). Kreitman and Aguadé (1986) developed a high-resolution approach to identify polymorphisms in genes in *Drosophila*. Their method involved digestion with frequently cutting (4-bp recognition sequence) restriction enzymes and electroblotting of fragments from sequencing gels. Thus, the yield of polymorphisms at a given locus can be increased, on average, 16-fold over that of traditional 6-bp restriction enzymes, because the recognition site of frequent cutters occurs at random every 256 nucleotides (Kreitman and Aguadé 1986). Alterations as small as single-nucleotide insertions and deletions can be detected using this technique. We used the PCR to apply this approach to a human anonymous DNA segment linked to cystic fibrosis (Dean et al. 1990*a*). Here we present a similar strategy to identify RFLPs in specific genes. Most expressed sequences could therefore be placed on genetic maps and be tested for linkage or association with genetic diseases.

Material and Methods

PCR

Each reaction contained 100 ng of DNA, 200 μM dNTPs, 1 unit *Taq* polymerase (Cetus or Digene), 1 μl of a 1 O.D./ml stock of each oligonucleotide in 10 mM Tris pH 8.8, 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelatin (Saiki et al. 1985). Reactions were performed in a volume of 25 μl. PCR products can be directly digested by adding 1 vol of a solution containing restriction enzyme and a twofold concentration of the recommended buffer. For single-stranded conformation polymorphisms (SSCP) analysis, PCR reactions were performed as above, except that 0.1 μl of ^{32}P-dCTP was added, the dNTPs were reduced to 70 μM, and the reaction volume was 10 μl. Two microliters of this reaction were mixed with 8 μl of 95% formamide, 5 mM NaOH, 0.1% bromophenol blue, and 0.1% xylene cyanol. The samples were heated to 95°C for 2 min and were cooled on ice, and 2 μl was loaded on 0.4-mm-thick 5% acrylamide gels in 1 × TBE. Electrophoresis was performed at 3 W for 16-20 h at room temperature, and the samples were dried onto 3 MM paper and were exposed to X-ray film for 3-48 h (Dean et al 1990*b;* Dean and Gerrard 1991). The use of 0.25 mM spermidine has been shown to increase the consistency of amplification of certain PCR primers (Straub and Bale 1990; Gerrard et al., submitted). In addition, incomplete denaturation can be overcome by diluting the PCR products prior to electrophoresis (Orita et al. 1989; data not shown). Primers employed for *KIT* (Yarden et al. 1987) were 5′TCTGAGCAGAATCAGTGTTTGGGTC and 5′CAGTAACTTTGTCAAACAGCATA. These primers amplified a 975-bp fragment that was digested with *Hae*III overnight at 37°C, and the fragments were separated by electrophoresis on 5% polyacrylamide gels. *IGF1R* oligonucleotides were 5′GAGACAGCTTCTCTGCAGTA and 5′TCCGGACACGAGGAATCAGC (Ullrich et al. 1986). The resulting fragment was digested with *Pvu*II and *Sau*96I for 2 h at 37°C and separated on 15% polyacrylamide gels.

Linkage Analysis

Linkage analysis was performed using version 3 of the CEPH data base. Files were converted into MAPMAKER format by using LNKTOMAP (provided by Ken Beutow) and were analyzed with the MAPMAKER program (Lander et al. 1987). After computing two-point LOD scores with each marker on the appropriate chromosome, the best order was calculated for the most closely linked loci.

Results

Use of the PCR to amplify DNA segments 500 bp or greater and digestion of the amplified product with frequently cutting restriction enzymes can identify polymorphisms that are not resolved by standard techniques (Dean et al. 1990*a*). Most mammalian genes contain both intervening sequences and 5′ untranslated regions that exhibit greater DNA variation than do coding domains, because variability of the latter segments is constrained by natural selection. DNA sequence information for most human loci is limited to cDNA transcripts, which, unfortunately for this analysis, do not identify the sequence position or length of intron sequences. However, the 3′ untrans-

Table 1

Polymorphisms in 3UTR

Gene (location) and Enzyme	Allele	Size (bp)	Frequency	*N*
KIT (4q11-12):				
*Hae*III	1	285	.83	70
	2	152 and 133	.17	
IGF1R (15q25-26):				
*Pvu*II + *Sau*96I.....	1	45	.75	100
	2	43	.25	

lated regions (3UTR) (generally greater than 300 bp) are usually reported, providing a DNA segment ideal to use in screening for polymorphisms. To directly test this idea, we amplified segments of the 3UTR from several genes and tested these regions for polymorphisms.

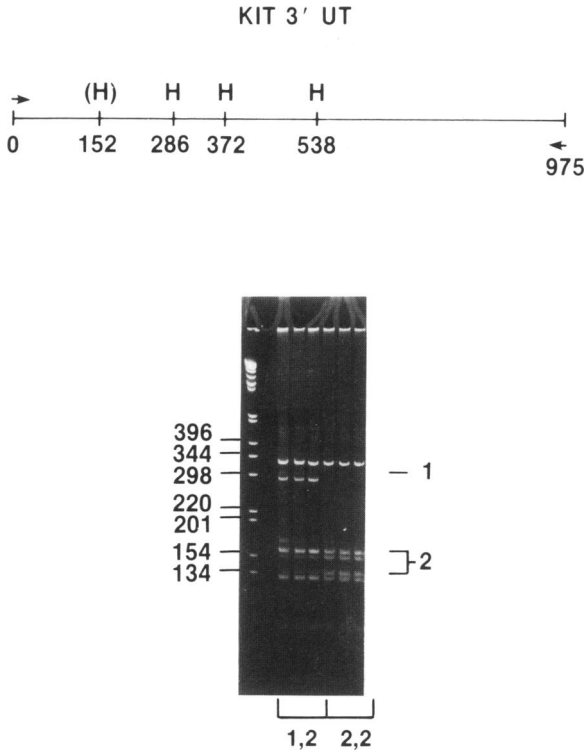

Figure 1 *Hae*III RFLP in 3UTR of *KIT* gene. *Top,* Map of *Hae*III restriction sites (H) in 3UTR of *KIT*. Arrows indicate the positions of the primers used for the PCR, and numbers denote the nucleotide positions of the *Hae*III sites. The (H) denotes the polymorphic site. *Bottom,* Ethidium bromide–stained 5% polyacrylamide gel of several unrelated individuals typed for *KIT* RFLP. The marker is the 1-kb ladder (BRL) with sizes shown in basepairs. 1, Allele 1. 2, Two fragments that constitute allele 2. Genotypes are shown below the lanes.

Table 1 presents two genes chosen for analysis: the proto-oncogene *KIT* and the insulin-like growth factor–1 receptor gene *IGF1R*. At the time this project was initiated, neither of these genes was known to be polymorphic. Primers from the *KIT*-gene 3UTR amplified a 975-bp segment consistent with the published sequence (Yarden et al. 1987). The segment was digested with a series of restriction enzymes with 4-bp recognition sequences, and the products resolved on 5% polyacrylamide gels. Fragments from *Hae*III digestion revealed a different pattern in several individuals, because of the presence of a polymorphic *Hae*III site (fig. 1, *top*). Examination of the map predicted from the sequence allowed the polymorphic site to be localized (fig. 1, *bottom*).

The *KIT* gene has been physically mapped to chromosome 4q11-12 by in situ hybridization (Yarden et al. 1987). To place the *KIT* gene on the linkage map, we typed all of the informative parents in the CEPH families for the *Hae*III polymorphism. The RFLP showed Mendelian segregation in all families analyzed, and the frequency of the alleles is given in table

Table 2

LOD Score Analysis of *KIT* and *IGFIR*

Gene and Marker	Location	θ_{max}	LOD Score
KIT:			
D4S35.....	4p11-q11	.02	6.98
D4S67.....	4p13-q13	.02	6.39
ATP1BL1	4p16-q23	.00	3.91
INP10.....	4q21	.20	1.20
IGF1R:			
D15S3.....	15	.00	5.10
D15S37.....	15	.42	.92

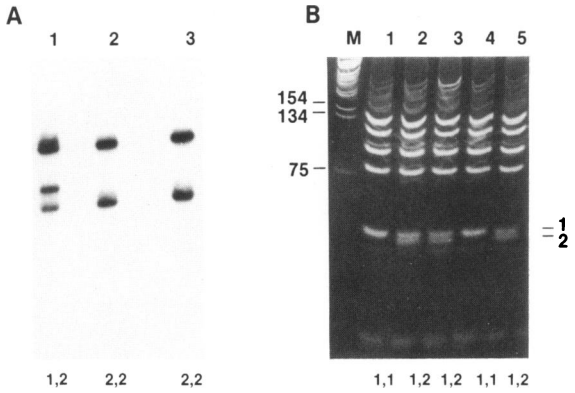

Figure 2 Polymorphism detected in *IGF1R* gene. *A*, Autoradiograph of SSCP gel resolving alleles of RFLP in *IFG1R* in three unrelated individuals. Allele 1 is resolved as a fragment of lower mobility; genotypes are shown below the lanes. *B*, Acrylamide gel resolving alleles after digestion with *Pvu*II and *Sau*96I.

1. A *Hin*dIII RFLP was recently described with a v-*kit* probe and was typed in these same families (Berdahl et al. 1988; Buetow et al. 1991). There was no recombination detected between the *Hae*III and *Hin*dIII RFLPs; moreover, the two markers are in strong linkage disequilibrium with a standardized linkage disequilibrium value (Δ) of .66 (χ^2 = 30.3). The *KIT* *Hae*III polymorphism showed significant linkage with several other chromosome 4 markers (table 2), and multipoint linkage analysis placed the gene between D4S35 and INP10, consistent with the physical location reported elsewhere (Yarden et al. 1987).

A similar polymorphism search of the *IGF1R* 3UTR revealed closely spaced, higher-molecular-weight bands after digestion with several enzymes, suggesting the presence of a small insertion/deletion polymorphism (data not shown). We have recently found that alleles that contain insertions as small as a single nucleotide can be efficiently resolved as SSCPs (Dean et al. 1990*b*). SSCPs are detected by fractionating denatured, radiolabeled DNA on nondenaturing gels (Orita et al. 1989). Figure 2 shows resolution of the *IGF1R* alleles on an SSCP gel. Individuals were also typed by double digestion with *Pvu*II and *Sau*96I (fig. 2*B*). Analysis of 14 informative CEPH families demonstrated that *IGF1R* is linked to the chromosome 15 marker D15S3 (θ = .00, LOD score = 7.5; table 2). D15S3 has been placed 23 cM distal to D15S37 and is the most distal marker on a linkage map of this region (Nakamura et al. 1988). This is consistent with the physical assignment of *IGF1R* to 15q25-26 (Ullrich et al. 1986) and provides an additional anchor point for the physical and genetic maps.

Table 3

Size of Human 3UTR

Locus	Size of 3UTR (bp)	Location	Spliced[a]
EGR1	1,380	5q23-q31	
EGR2	1,110	10q21	
CTLA4	1,132	2q33	No
G_3G_1	1,785		
CRYG	40	2q33-35	No
Gxα	2,050		
CCK	566	3pter-p21	No
GNAI1	512	7q21-22	
GLI2	140	2	No
GNAI2	3		Yes
GNAI3	1,100	1	
F11	180	4q35	No
ICAM-2	121		
F12	140	5q33-qter	No
CDR	432	Xq27	
PIM	1,330	6p21	No
Leukosiali	994	16	
TNFA	790	6p21.3	No
APP	1,108	21q21	
L2G25B	410		
4-1BB	1,435		
TCRA	550	14q211	
INSR	1,045	19p13	
VAV	265		
GF-1	500		
SYN	455		
MBP	1,603	18q22-ter	
Lipocurtin	286	9q11-22	
Glucocerebos	110	1q	
ANP	303		
TBG	295	Xq21-22	
PLP	540	Xq21-22	
CHRND	280	2q33-ter	No
IL2R	525	4q26-27	
LYN	480	8q13-ter	
IGF	280	12q23	
INHA	280	2q33-ter	
INHBA	130	7p15-p13	
INHBB	900	2cen-q13	
PDGFR	1,130	5q31-33	
PKCB	830	16p11.2-12	
21-OHase	490		
IL3	415	5q23-31	
CFAG	66	1	
CD1	580	1q22-23	

NOTE.—Genes chosen at random were examined for the length of the 3UTR. References for each locus can be found in the work of Kidd et al. (1989).

[a] Entries indicate whether the 3UTR is disrupted by an intron; blank indicates that the answer is unknown.

Discussion

We describe here an empirical strategy for identifying RFLPs for virtually any coding-gene sequence. The combination of PCR amplification of untranslated domain segments and high-resolution methods for visualizing RFLPs should allow the development of useful polymorphism for many of the 2,000 coding genes that are on the human gene map but not yet on the linkage maps. An examination of 45 randomly chosen genes suggests that most genes have a 3UTR that is greater than 300 bp (table 3). This approach will allow genes to be (*a*) more reliably placed on the genetic maps of human chromosomes as well as (*b*) tested for linkage and/or association with disease. Polymorphisms can be detected either by digestion with enzymes or by the identification of SSCPs. For regions of low variation, the SSCP method should be more efficient. SSCP can potentially detect all variations, and we have identified point mutations by using SSCP that cannot be detected with restriction enzymes.

The degree of polymorphism within the 3UTR is not well known. Several sequences of 3UTRs from homologues have been described. In the *SRC* gene there is no significant homology between human 3UTR and chicken 3UTR (Anderson et al. 1985). For the *PKCG* gene the overall similarity with the bovine gene is 50%; however, the extent of identity within the 3UTR of these genes varies, being 10%–90% (Coussens et al. 1986). The *PIM1* gene 3 UTR is 72% conserved between mouse and human, in contrast to the 94% identity in the coding region (Zakut-Houri et al. 1987). An exceptional case is the *IL3* gene, for which the 3UTR of the mouse gene was used to clone the human homologue (Dorssers et al. 1987). The 3UTR contains sequences for poly-A addition and may also be involved in termination, transport, and stability of mRNAs (Jackson and Standart 1990). However, the insertion of repetitive sequences and the reduced conservation of these regions suggest that there is little evolutionary constraint on most of the domain.

Other regions of genes may also prove to be useful in identifying RFLPs. For some genes the location and size of the intervening sequences are known, allowing these regions to be amplified. Introns not only vary rapidly but in some cases contain highly polymorphic sequences, such as VNTRs and minisatellites (Furutani et al. 1986; M. Dean, unpublished data). Many introns also contain *Alu*-family repetitive sequences, many of which may be polymorphic (Epstein et al. 1990).

At least three distinct types of mapping endeavors are currently used to characterize the human genome; these are the genetic, physical, and comparative approaches. The order of genes on murine chromosomes provides useful clues to the order of human sequences. Genes provide the anchor points that connect the linkage and physical maps of many of the chromosomes. The further identification of RFLPs in expressed sequences can only serve to increase the usefulness of all three types of mapping approaches. The 3UTRs of genes provide convenient sequence-tagged sites (STS) that have been proposed as being anchor points of chromosome maps (Olson et al. 1989). Polymorphic STS will be important for connecting the physical and genetic maps.

Acknowledgments

We thank Anjanette Perry for technical assistance, and we thank Krista Hampsch and the ASCL for computer assistance. This project has been funded in part with Federal funds from the Department of Health and Human Services under contract number NO1-CO-74102 with Program Resources, Incorporated/DynCorp.

References

Anderson SK, Gibbs CP, Tanaka A, Kung H-J, Fujita DJ (1985) Human cellular *src* gene: nucleotide sequence and derived amino acid sequence of the region coding for the carboxy-terminal two-thirds of $pp60^{c\text{-}src}$. Mol Cell Biol 5: 1122–1129

Berdahl LD, Murray JC, Besmer P (1988) A *Hin*dIII RFLP demonstrated for the kit oncogene on chromosome 4. Nucleic Acids Res 16:4760

Buetow KH, Shiang R, Yang P, Nakamura Y, Lathrop GM, White R, Wasmuth JJ, et al (1991) A detailed multipoint map of human chromosome 4 provides evidence for linkage heterogeneity and position-specific recombination rates. Am J Hum Genet 48:911–925

Coussens L, Parker PJ, Rhee L, Yang-Feng TL, Chen E, Waterfield MD, Francke U, et al (1986) Multiple, distinct forms of bovine and human protein kinase C suggest diversity in cellular signaling pathways. Science 233:859–866

Daussett J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R (1990) Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6:676–677

Dean M, Drumm JL, Stewart C, Gerrard B, Perry A, Hidaka N, Cole JL, et al (1990*a*) Approaches to localizing disease

genes as applied to cystic fibrosis. Nucleic Acids Res 18: 345–350

Dean M, Gerrard B (1991) Helpful hints for the detection of single-stranded conformation polymorphisms. BioTechniques 10:332–333

Dean M, White MB, Amos J, Gerrard B, Stewart C, Khaw K, Leppert M (1990*b*) Multiple mutations in highly conserved residues are found in mildly affected cystic fibrosis patients. Cell 61:863–870

Dorssers L, Burger H, Bot F, Delwel R, Geurts van Kessel, HM, Löwenberg B, Wagemaker G (1987) Characterization of a human multilineage-colony-stimulating factor cDNA clone identified by a conserved noncoding sequence in mouse interleukin-3. Gene 55:115–124

Epstein N, Nahor O, Silver J (1990) The 3′ ends of *alu* are highly polymorphic. Nucleic Acids Res 18:4634

Furutani Y, Notake M, Kukui T, Ohue M, Nomura H, Yamada M, Nakamura S (1986) Complete nucleotide sequence of the gene for human interleukin 1 alpha. Nucleic Acids Res 14:7897–7914

Gerrard BC, Lucas-Derse S, Dean M. Increased consistency and efficiency of PCR reactions by the addition of spermidine (submitted)

Jackson RL, Standart N (1990) Do the poly(A) tail and 3′ untranslated region control mRNA translation? Cell 62: 15–24

Kidd KK, Bowcock AM, Schmidtke J, Track RK, Ricciuti F, Hutchings G, Bale A, et al (1989) Report of the DNA committee and catalogs of cloned and mapped genes and DNA polymorphisms. Cytogenet Cell Genet 51:622–947

Kreitman M, Aguadé M (1986) Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hybridization of four-nucleotide-recognition restriction enzyme digests. Proc Natl Acad Sci USA 83:3562–3564

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln DE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181

Nakamura Y, Lathrop M, O'Connell P, Leppert M, Lalouel JM, White R (1988) A mapped set of DNA markers for human chromosome 15. Genomics 3:342–346

O'Brien SJ, Graves JAM (1990) Report of the Committee on Comparative Gene Mapping. Human Gene Mapping 10.5: International Workshop on Comparative Gene Mapping. Cytogenet Cell Genet 55:406–433

Olson M, Hood L, Cantor C, Botstein D (1989) A common language for physical mapping of the human genome. Science 245:1434–1435

Orita M, Suzuki Y, Sekiya T, Hayashi K (1989) Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. Genomics 5: 874–879

Saiki R, Scharf S, Faloona F, Mullis K, Horn GT, Erlich MA, Arnheimn N (1985) Enzymatic amplification of beta-globin genomic sequence and restriction site analysis for diagnosis of sickle cell anemia. Science 230:1350–1354

Straub TD, Bale AE (1990) Spermidine overcomes the effect of Taq polymerase inhibitor of PCR: implications for DNA diagnostics with multiplex reactions. Am J Hum Genet 47:A237

Ullrich A, Gray A, Tam AW, Yang-Feng T, Tsubokawa M, Collins C, Henzel W, et al (1986) Insulin-like factor I receptor primary structure: comparison with insulin receptor suggests structural determinants that define functional specificity. EMBO J 5:2503–2512

Yarden Y, Kuang W-J, Yang-Feng T, Coussens L, Munemitsu S, Dull TJ, Chen E, et al (1987) Human proto oncogene c-*kit:* a new cell surface receptor tyrosine kinase for an unidentified ligand. EMBO J 6:3341–3351

Zakut-Houri R, Hazum S, Givol D, Telerman A (1987) The cDNA sequence and gene analysis of the human *pim* oncogene. Gene 54:105–111